



dCache, some details

Seminar at the Laboratori Nazionali di Legnaro, INFN

Patrick Fuhrmann



Content

- The project structure
 - Partners and people
 - Our funding
 - Sustainability/Networking
- Deployments
 - WLCG overall
 - Big
 - Wide
 - Super Mini
- Customers Relation
 - Deployment Channels
 - User Support channels
- Technology (Design -> consequence)
 - Location independent Services
 - Metadata , Data separation (Multi Tier 3)
 - gPlazma (IdP)
- Work in progress
 - For WLCG
 - For Photon Science
 - Cloud software and service



Cheat Sheet



Cheat Sheet

- dCache.org is an international collaboration, developing and distributing storage software (dCache)
- dCache is in production in about 60 places around the world and stores (roughly) about 120 Pbytes in total for WLCG.
- dCache supports different storage media, like disk, SSD and tape and provides mechanisms for manual and automated internal and external replication and transitions.
- dCache storage can be accessed via standard protocols like WebDAV, NFS, and gridFTP and proprietary protocols like dCap and xrootd.
- dCache supports a variety of authentication and mapping mechanisms, e.g. Kerberos, X509, User/Password, LDAP, NIS, NSSWITCH.



Project Structure

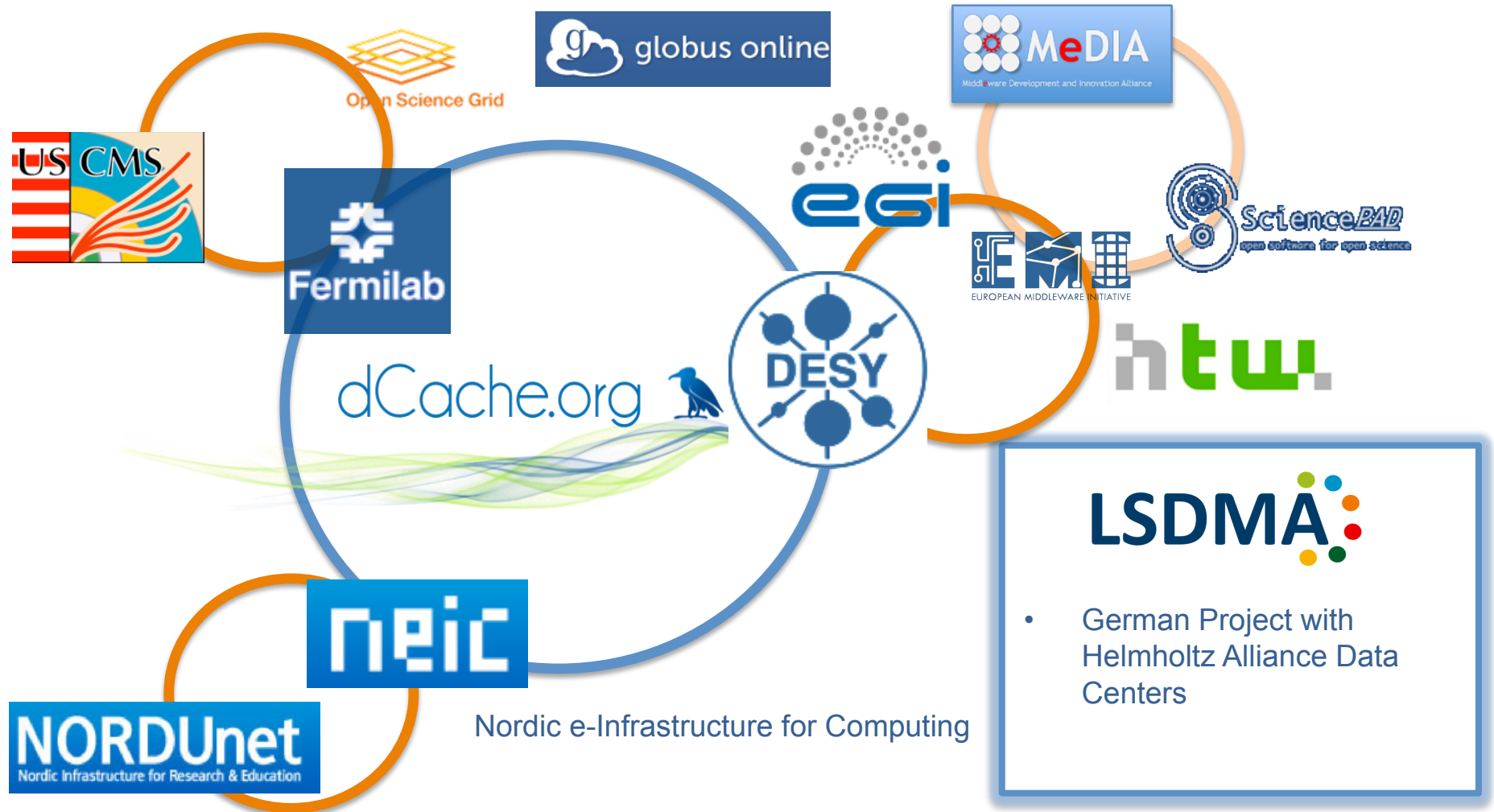
The dCache partners and team

dCache.org



dCache.org@DESY will start hiring in a couple of weeks.

dCache partners bridging national projects and activities.





Data Lifecycle Labs (Customers)

- Energy
 - smart grids, battery research, fusion research
- Earth and Environment
- Health
- Key Technologies
 - synchrotron radiation, nanoscopy, high throughput microscopes, electron-microscope imaging techniques
- Structure of Matter

Data Service Integration Team

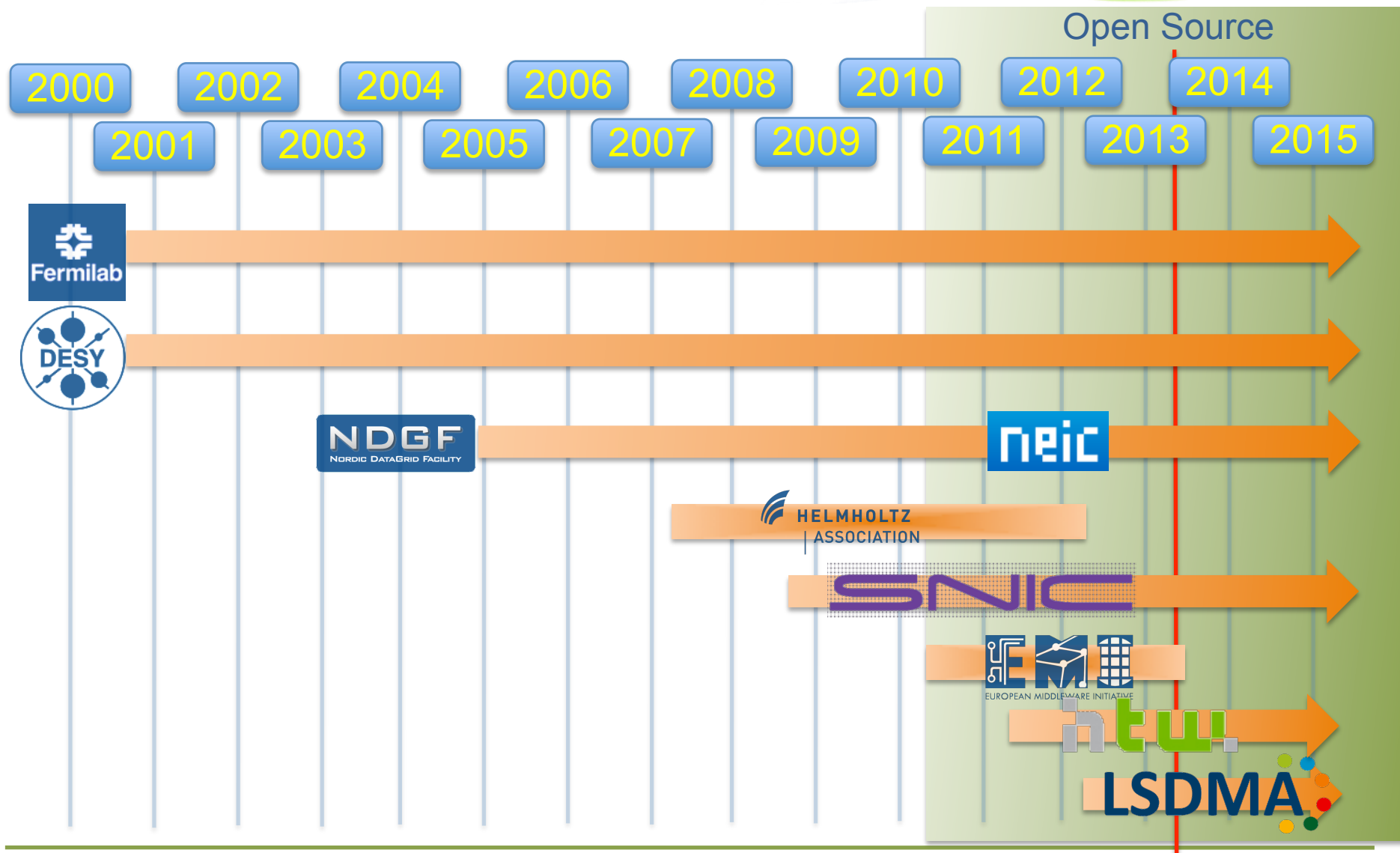
dCache.org



- **Federated Identity**
- Federated Data Access
- Metadata Management
- Archiving

Funding and Partners

dCache project timeline



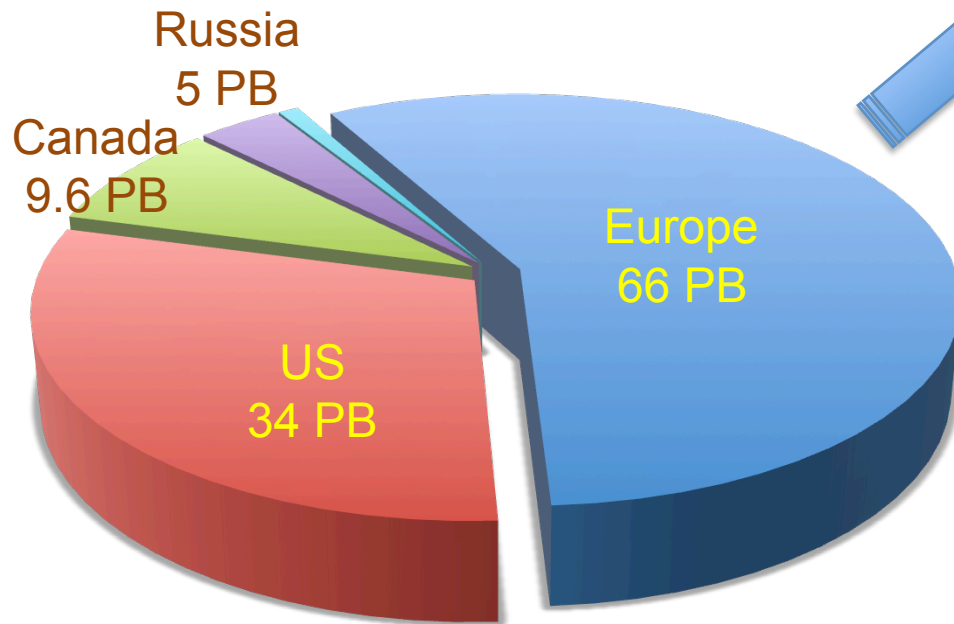
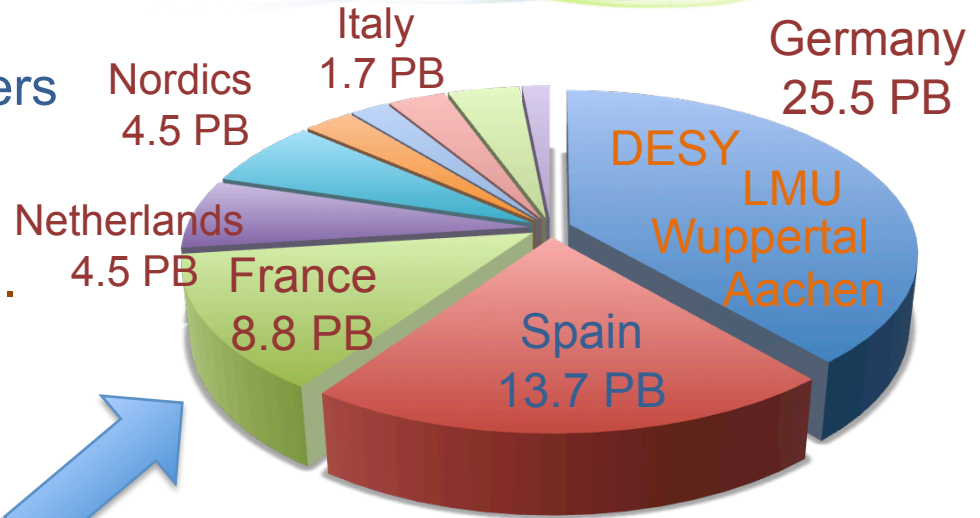


Deployments

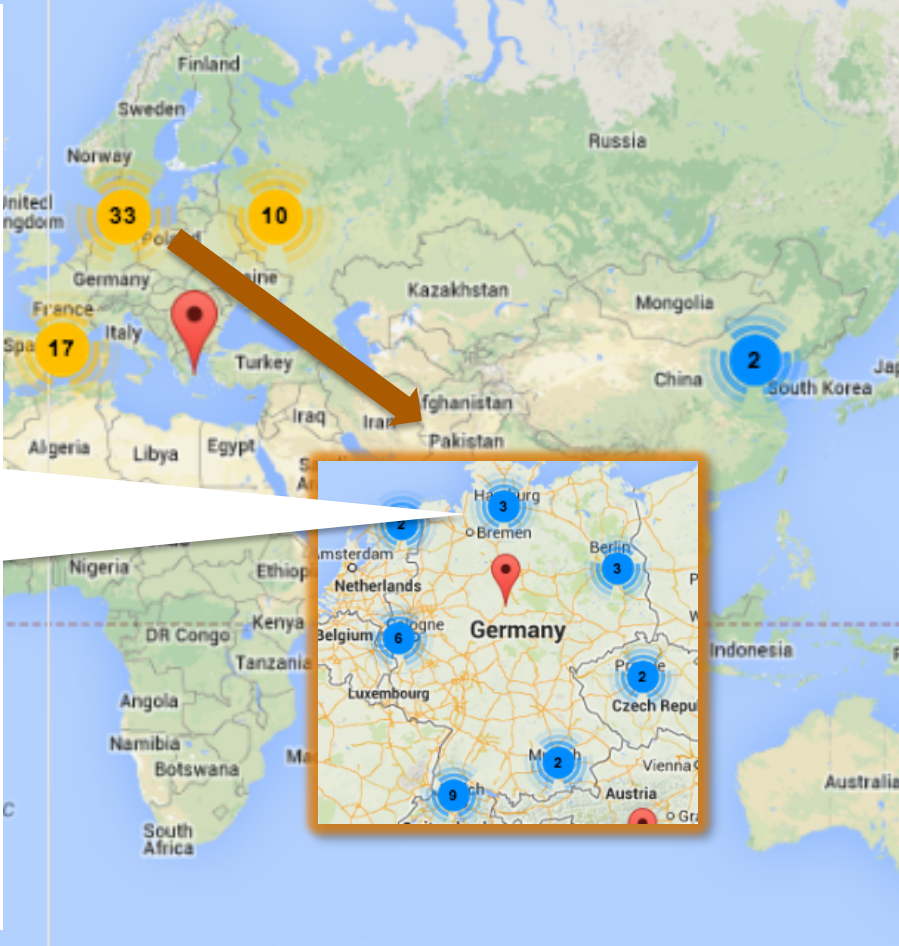
dCache storage for WLCG



- About 115 PBytes just for WLCG
- In 8(+2) out of 11(+3) Tier 1 centers
- And about 60 Tier 2's, which is
- about 1/2 of the entire WLCG data.



Tigrans new dCache world map dCache.org



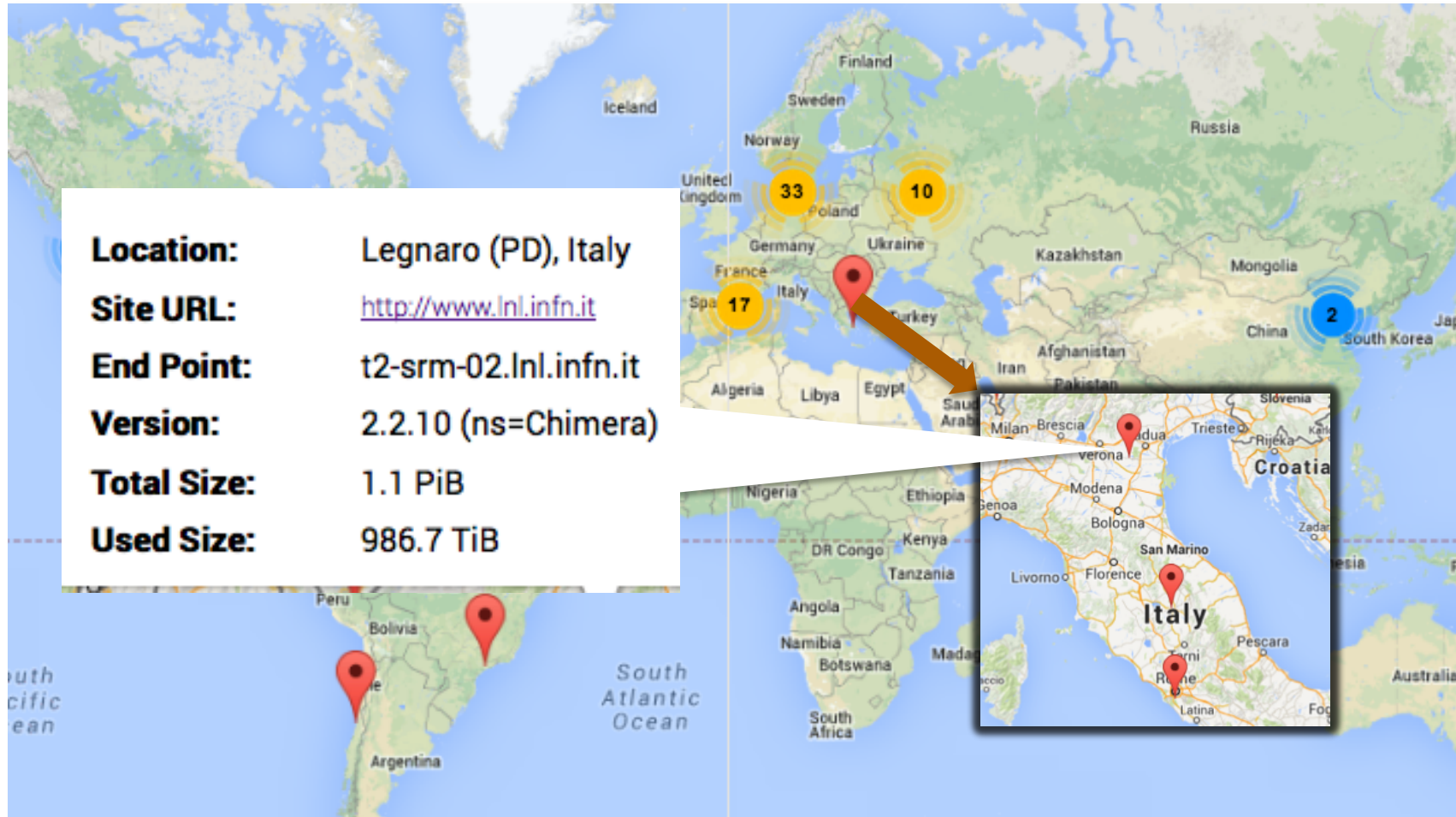
The world map displays various dCache sites across different continents. Callouts are shown for several locations: 33 in the UK, 10 in Poland, 17 in Spain, 2 in China, and 2 in South Korea. A red pin marks Hamburg, Germany, with a callout box providing details for three different dCache instances at that location. A brown arrow points from the Hamburg location on the world map to the callout box.

Instance	Location	Site URL	End Point	Version	Total Size	Used Size
1	Hamburg, Germany	http://grid.desy.de/	dcache-se-desy.desy.de	2.6.5 (ns=Chimera)	718.2 TiB	246.6 TiB
2	Hamburg, Germany	http://grid.desy.de/	dcache-se-cms.desy.de	2.6.6 (ns=Chimera)	3.9 PiB	3.6 PiB
3	Hamburg, Germany	http://grid.desy.de/	dcache-se-atlas.desy.de	1.9.12-12 (ns=Chimera)	2.6 PiB	2.0 PiB

Available at dCache.org

Tigrans new dCache world map

dCache.org



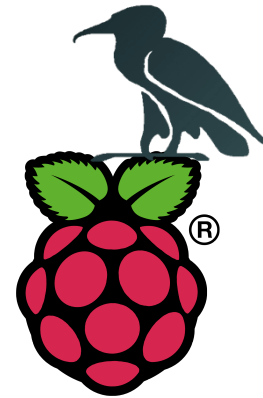
Available at dCache.org



Interesting installations

The raspberry dCache

dCache.org

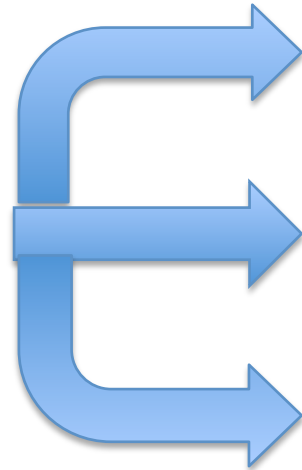


700 MHz ARM
512 MB Memory
2 * USB 2
100 MB Ethernet



Customer Relations

Deployment Channels



dCache.ORG / Web Pages

 UMD

Targeting: EPEL

- support reports, distribution
- German dCache operation annual
- EGI.eu: UMD.
- Weekly
- 2 dCac

The first Asian Pacific dCache Workshop

17 March 2013 - Taipei

Main Topics

- dCache Installation & Configuration
- NFS4.1/pNFS
- HTTP/WebDAV
- Security
- Hardware Life Cycle
- Tertiary Storage Access
- dCache Features
- Master Classes



) for all bug
p. Tickets are

German
ring and daily
tutorial of the

ges taken from



Technology and design

Design Patterns and consequences

(stolen from a dCache tutorial)



Design #1 Service Modules & Message Passing

Design #2 Namespace – Physical Storage separation

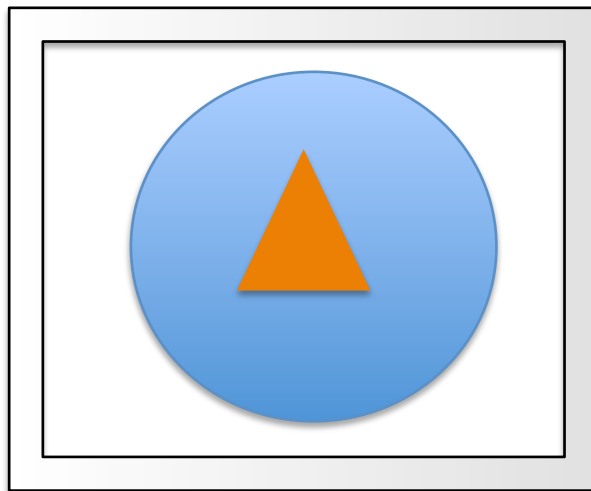
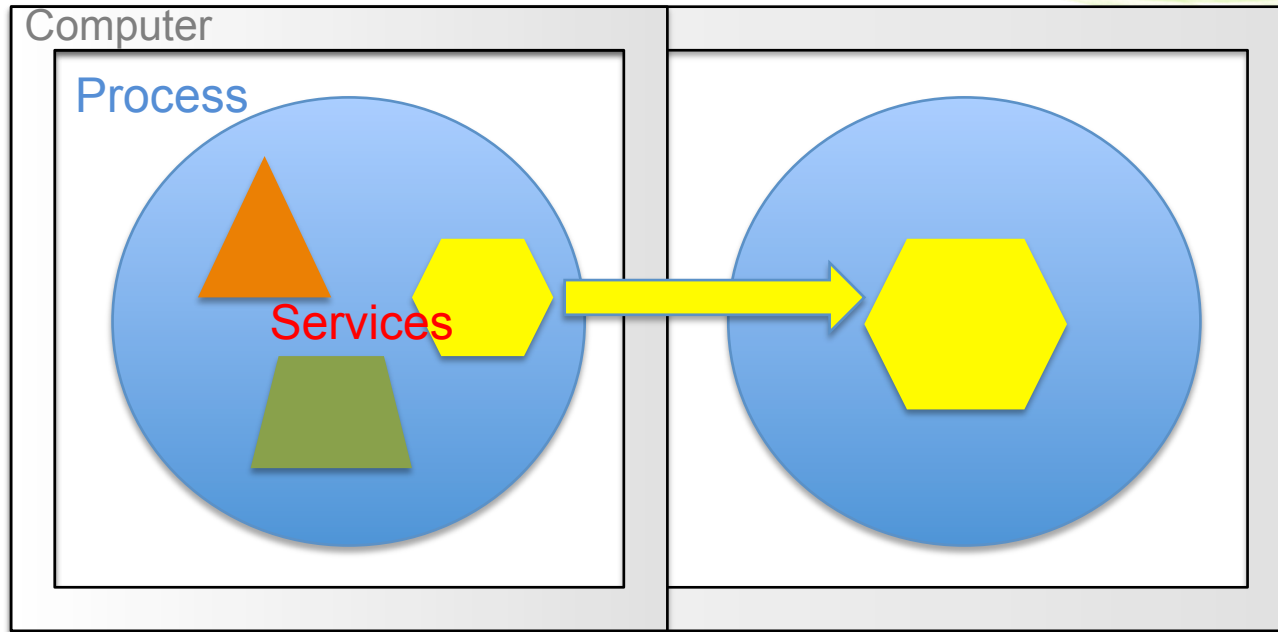
Design #3 Services allow plug-ins



Design #1

Service Modules & Message Passing

Scale-out Design

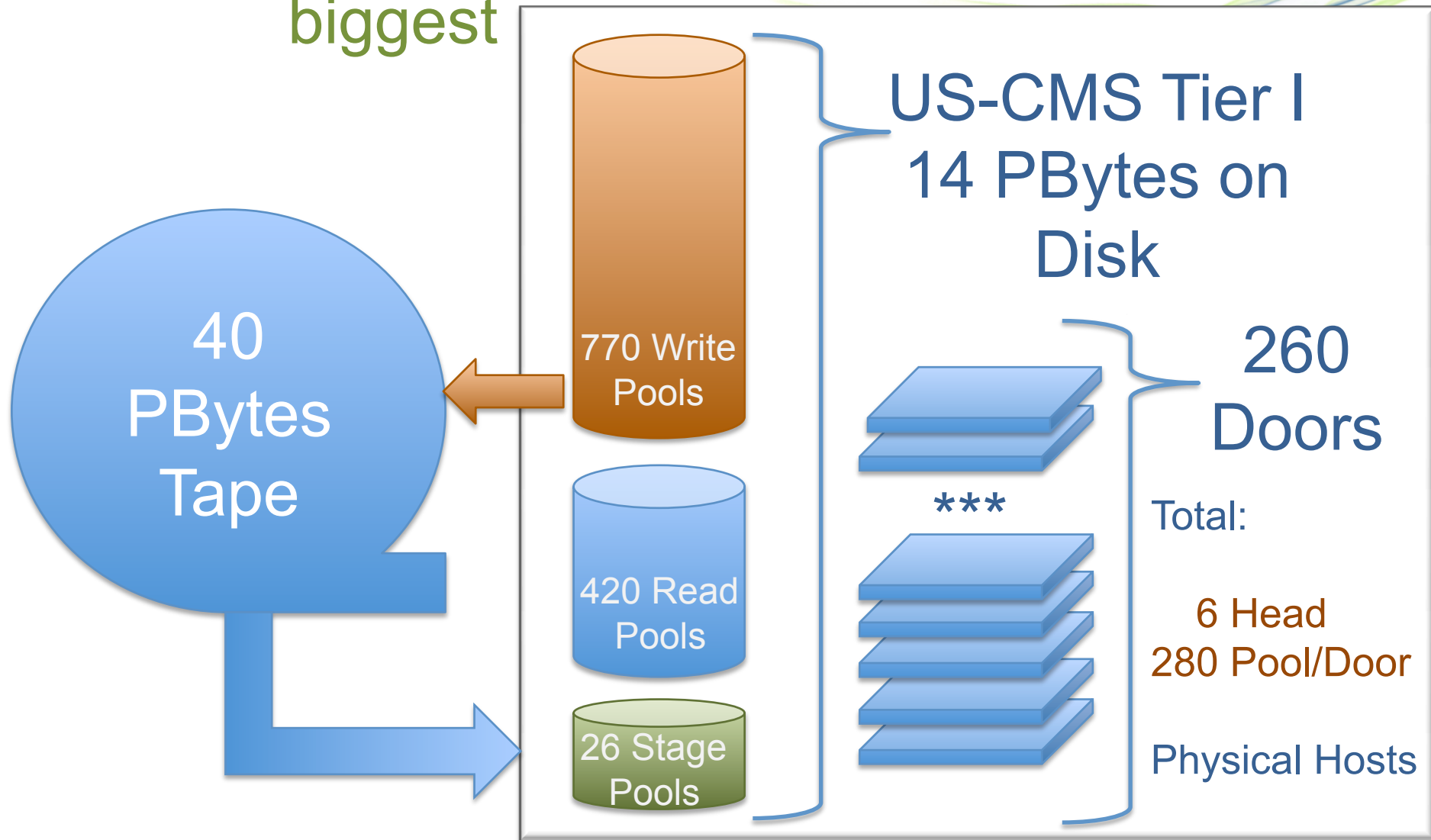


- Services are location independent.
- Services communicate via messages.



Resulting in Fits all sizes

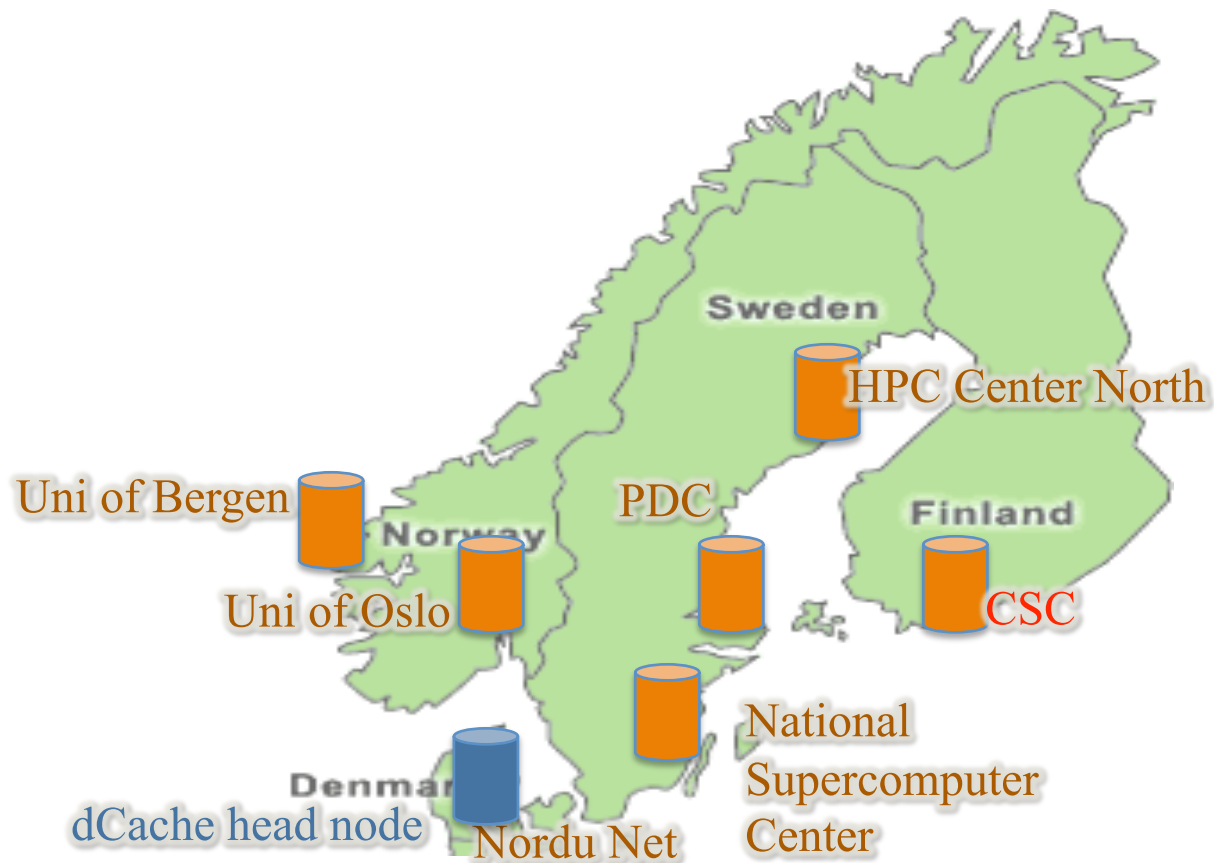
Starting with possibly the biggest



Information provided by Catalin Dumitrescu and Dmitry Litvintsev

To certainly the
most widespread

dCache.org



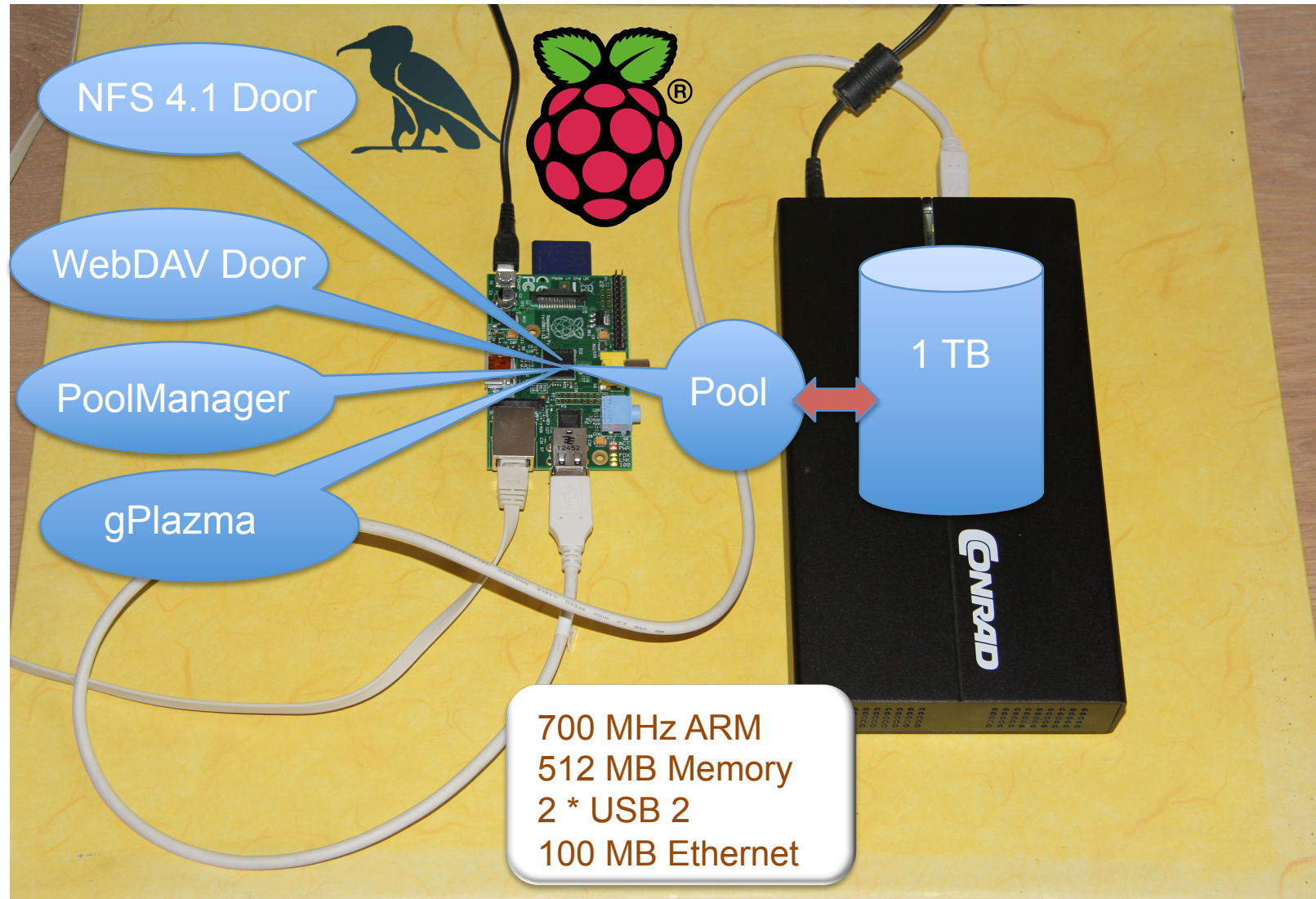
4 Countries

One dCache

Slide stolen from Mattias Wadenstein, NDGF

To very likely the smallest

One Machine – One Process



Design and consequence

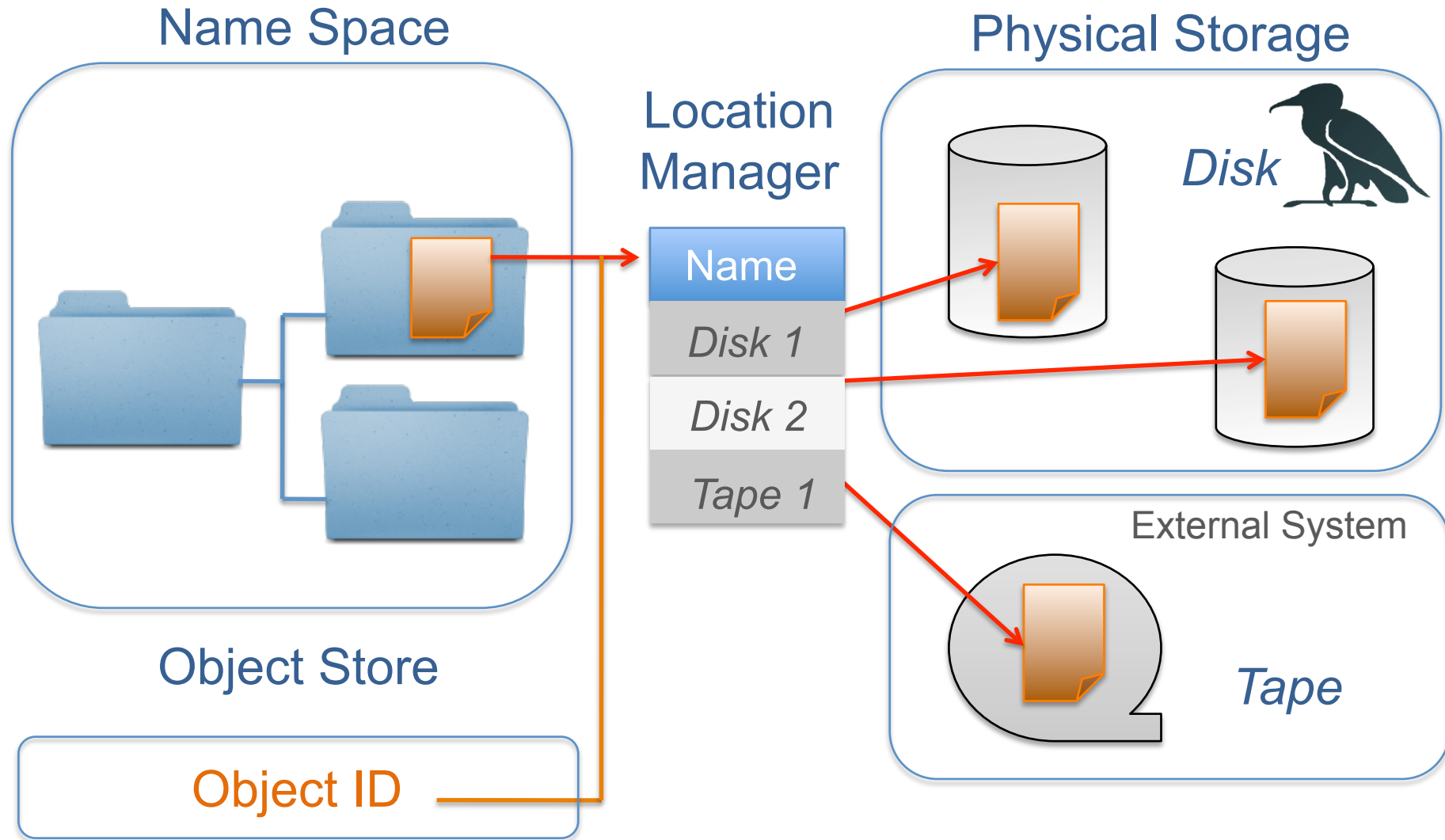
dCache.org



Design #2 Namespace – Physical Storage separation

Design

Namespace – Storage separation





Resulting in Replica Management

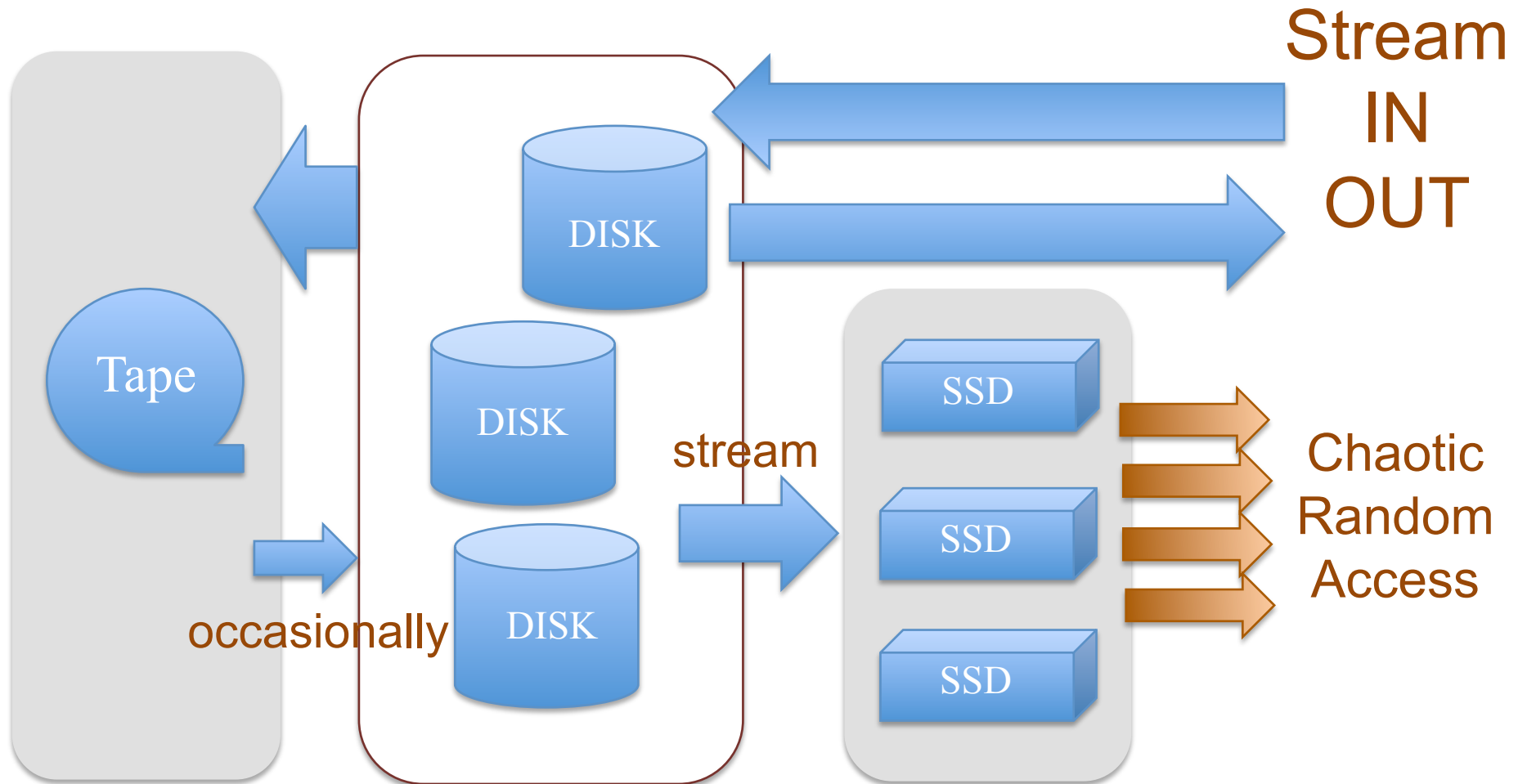
Replica Management

- Hot Spot detection
 - Files are copied from 'hot' to 'cold' pools
- Multi Media Support
 - File location is based on access profile and storage media type/properties
 - Fast streaming from spinning disks
 - Fast random I/O from SSD's
- Migration Module(s)
 - Files can be manually/automatically moved or copied between pools.
 - Rebalancing of data after adding new (empty) pools.
 - Decommission pools.
- Resilient Manager
 - Keeps max 'n' min 'm' copies of a file on different machines.
 - System resilient against pool failures.
- Tertiary System connectivity (Tape systems)
 - Data is automatically migrating to tape.
 - Data is restored from tape if no longer on disk

In preparation : Multi Tier Storage dCache.org



Analysis



Design and consequence

(stolen from a dCache tutorial)

dCache.org



Design #3

Services allow plug-ins

Plug-in Facility



Standard File Access Protocols

http(s)
WebDav

NFS 4.1

gsiFtp

Storage Management

SRM



Common Security Layer

Authentication : Kerberos, X509, Password

Unified ID management

Authorization : ACL's for File system and storage control (SRM)

Common Name Service Layer

Extended Names Service Queries (SQL)

“multi-media” storage layer

DISK

DISK

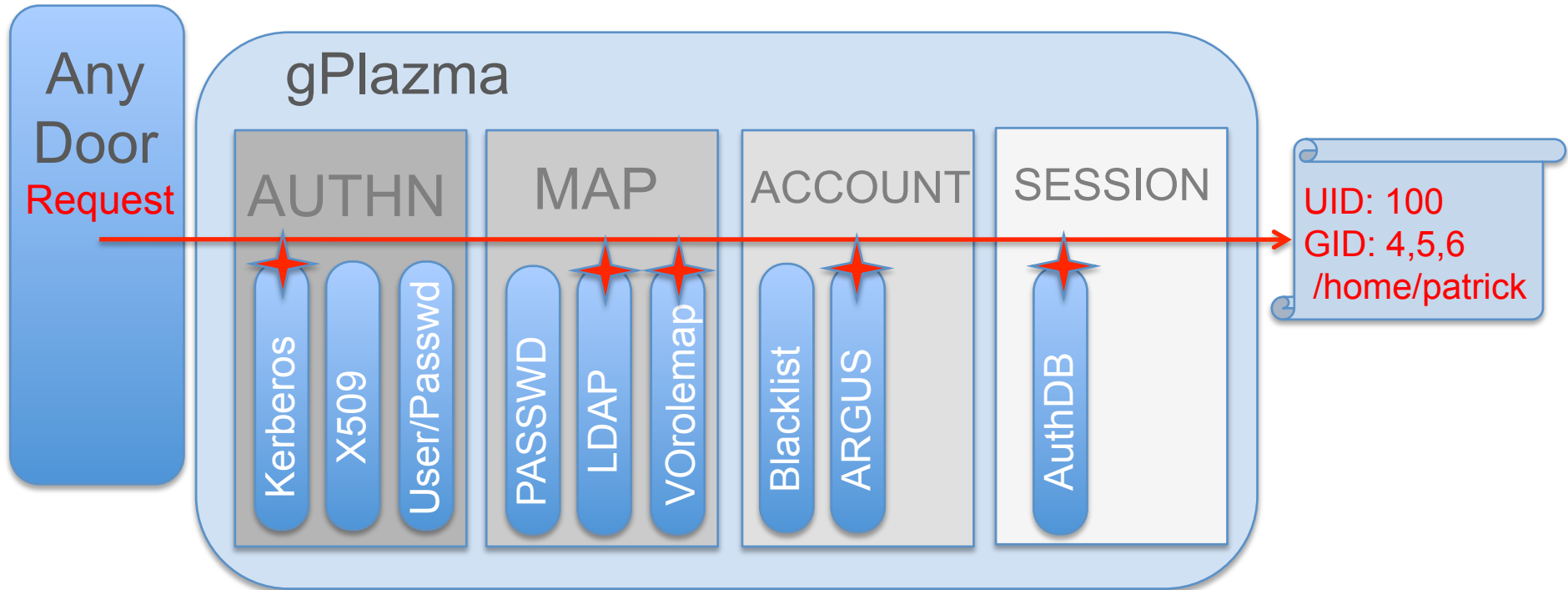
SSD

SSD

Tape

gPlazma (AAI) pluggin

Design stolen from Paul



Out of the box plug-ins

- Kerberos
- X509 Certs and proxies
- User / Password
- NIS/LDAP
- NSSWITCH
- GridMapFile
- VO Rolemap
- Argus
- Local Blacklist

Consequence of #3

- Authentication, mapping and user attributes are handled separately and in independent pluggins. So any reasonable combination is possible.
- Data access protocols and gPlazma are orthogonal. So the same mapping and user handling can be applied to any protocol.
- Easy to add a new auth/mapping pluggin to handle local infrastructure. (e.g. SARA added their own LDAP pluggin)



Work in progress

WLCG

Photon Science

General Scientific Groups



- Quick reminder:
 - pNFS allows GRID storage elements (e.g. dCache and DPM) to be mounted like regular disk systems.
 - But provides scaling by letting the client directly exchanging data with the individual storage node.
 - Photon Science and BELLE (1&2) are already accessing their data via NFS at DESYs dCaches.
- As SL6 is now ready for WLCG, NFS 4.1/pNFS clients are available on work group servers and worker nodes.
- CMS and ALTAS dCache at DESY have been upgraded, supporting latest NFS4.1/pNFS server.
- DESY is evaluating NFS for CMS (thanks to Christoph Wissing and DOT Team), starting with the “National Analysis Facility”, followed by GRID worker nodes.
- Next step will be evaluations at FERMIlab.

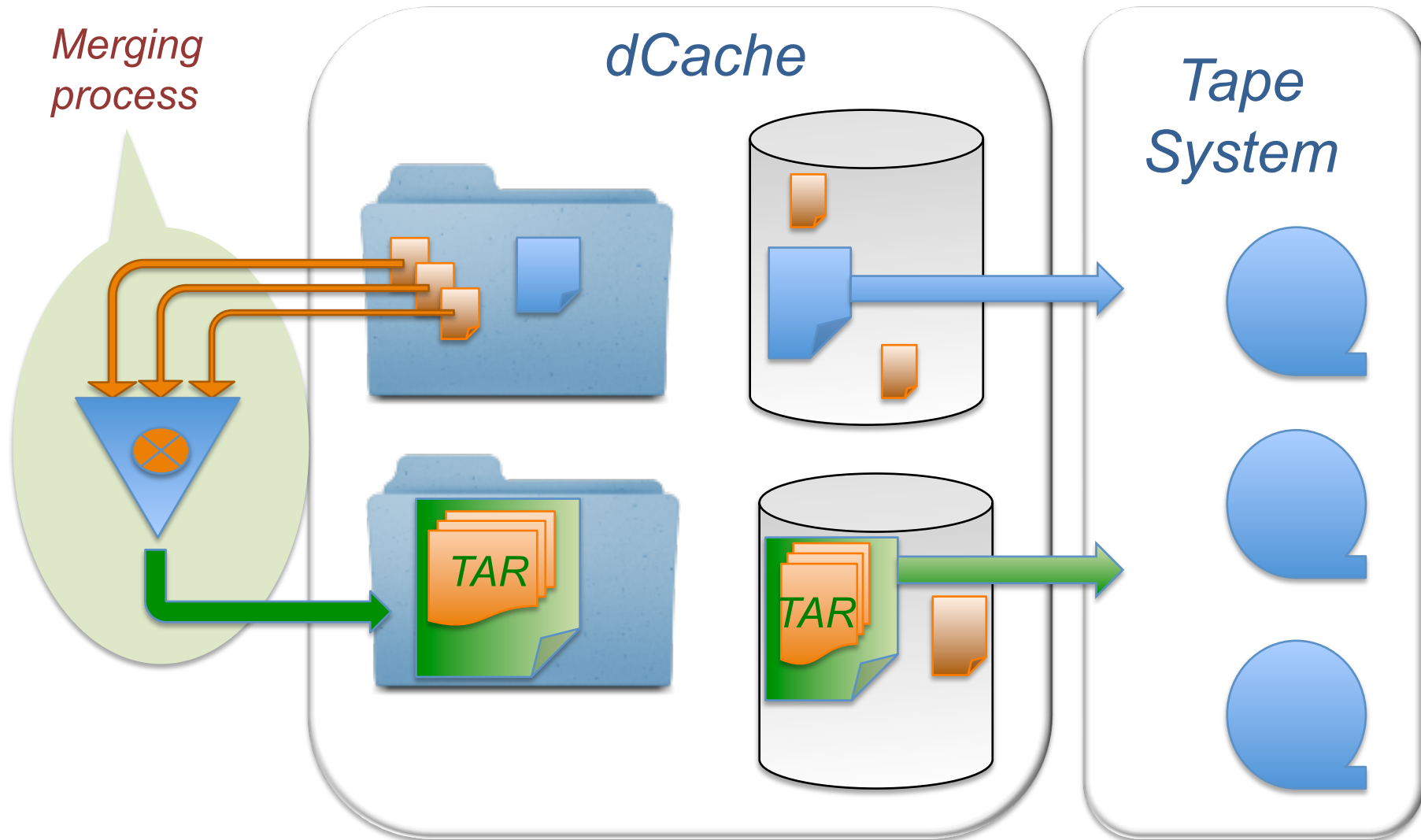
Other WLCG stuff

- Multi Tier storage (Tape, Disk, SSD)
- Adding hooks for the CMS and Atlas data xRootd federation
- CMS Disk / Tape separation

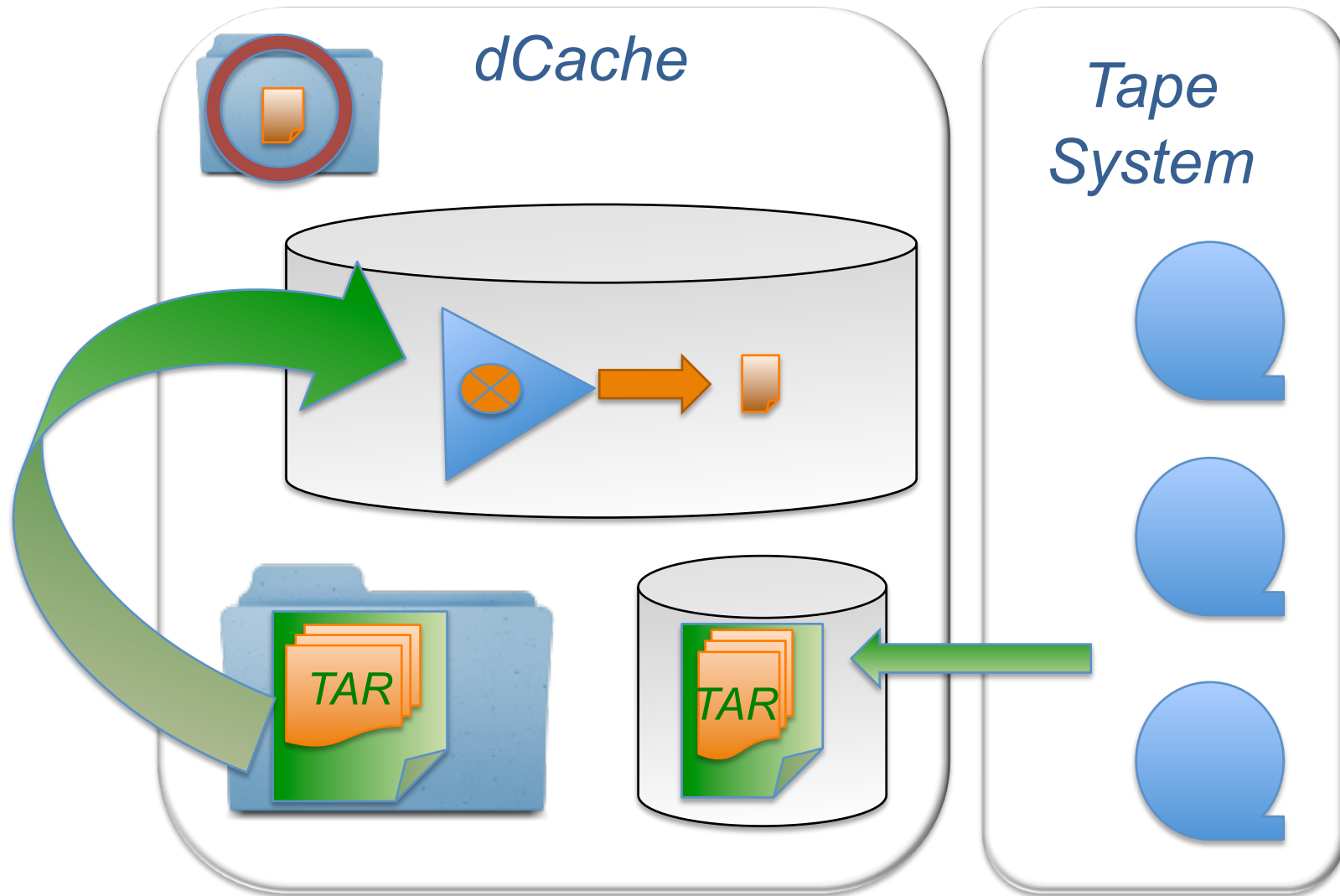


- Small File Support
 - For now only a service at DESY, but will be integrated into the code when sufficiently tested.
 - See next slide
- Fast Single Stream Injection
 - Right now we can achieve 30 Hz, single stream
- Support of HDF5, Nexus file formats
 - That's actually 'read/modify/write' for dCache files 😊

Support for small files



Extracting

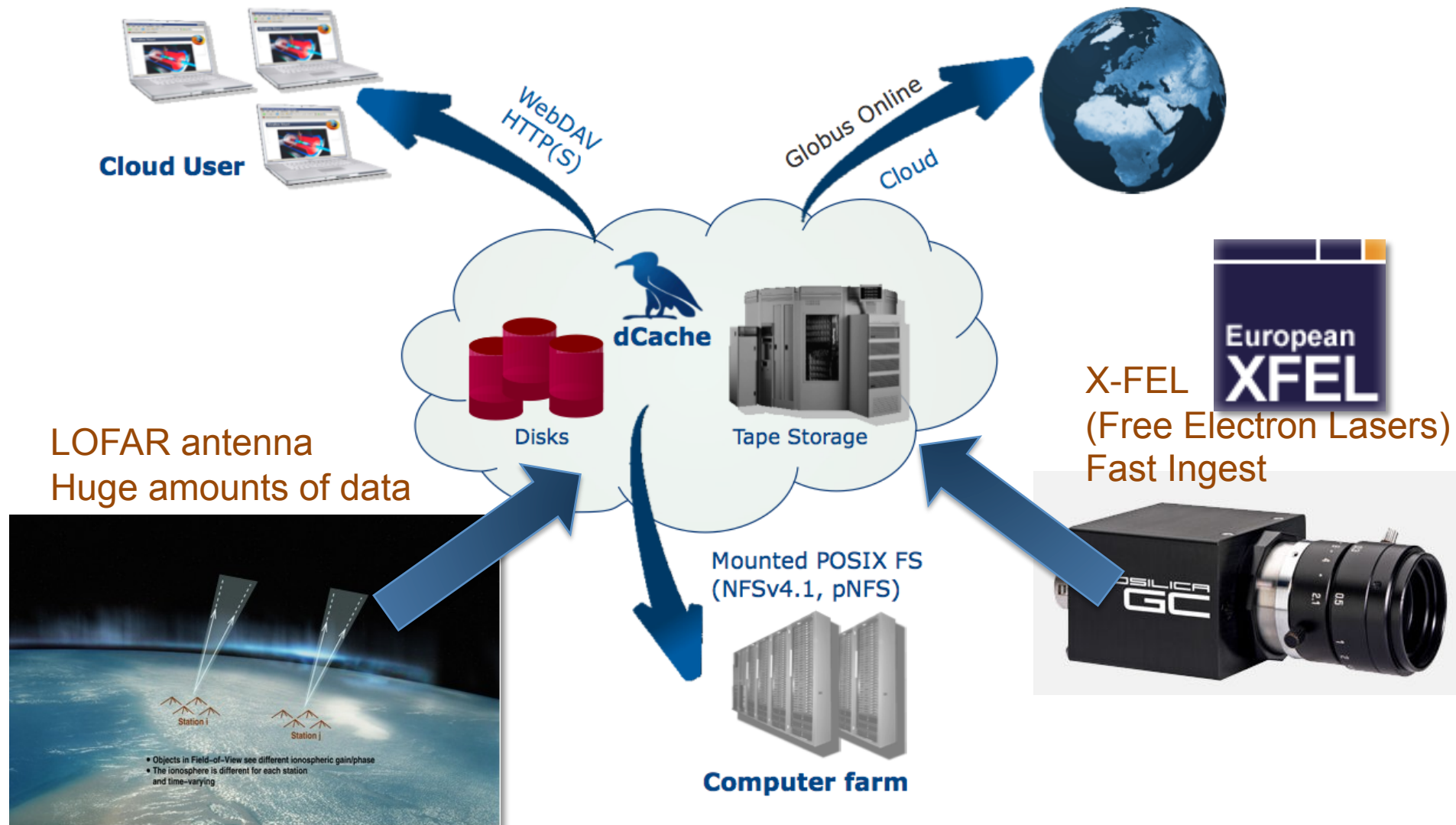




And now for something completely different

The scientific storage cloud

Scientific Storage Cloud



Scientific Storage Cloud Requirements

dCache.org



- Data can be accessed by a variety of protocols
 - Globus-online transfers via **gridFTP**
 - FTS Transfers for WLCG via gridFTP or **WebDAV**
 - Private upload and download via **WebDAV**
 - Public anonymous access via plain **http(s)**
 - Direct fast access from worker-nodes via **NFS4.1/pNFS** (just a mount like GPFS or Lustre but with standards)
- Individuals are authenticated by different mechanisms
 - X509 certificates or proxies
 - Username/password
 - SAML assertions (from IdP)
 - Kerberos tokens



- In collaboration with the HTW Berlin, dCache.org makes a storage cloud available for students.
- They get unlimited storage for free and their masters or bachelor degree.
- We get:
 - Cloud CDMI Implementation
 - Knowledge what young people need from cloud.
 - Client software on mobile devices.
- Currently 3 students are working with us.

LSDMA Federated Identity

(Just one example)

- Organizing German sites/universities to provide IdP.
- Integrating those IdPs into the German DFN Online CA. (Hopefully including Umbrella/PanData).
- Goal: Easy access for all scientists, registered at any IdP, to access scientific resources (e.g. dCache) w/o handling X509 Certificates.



In summary

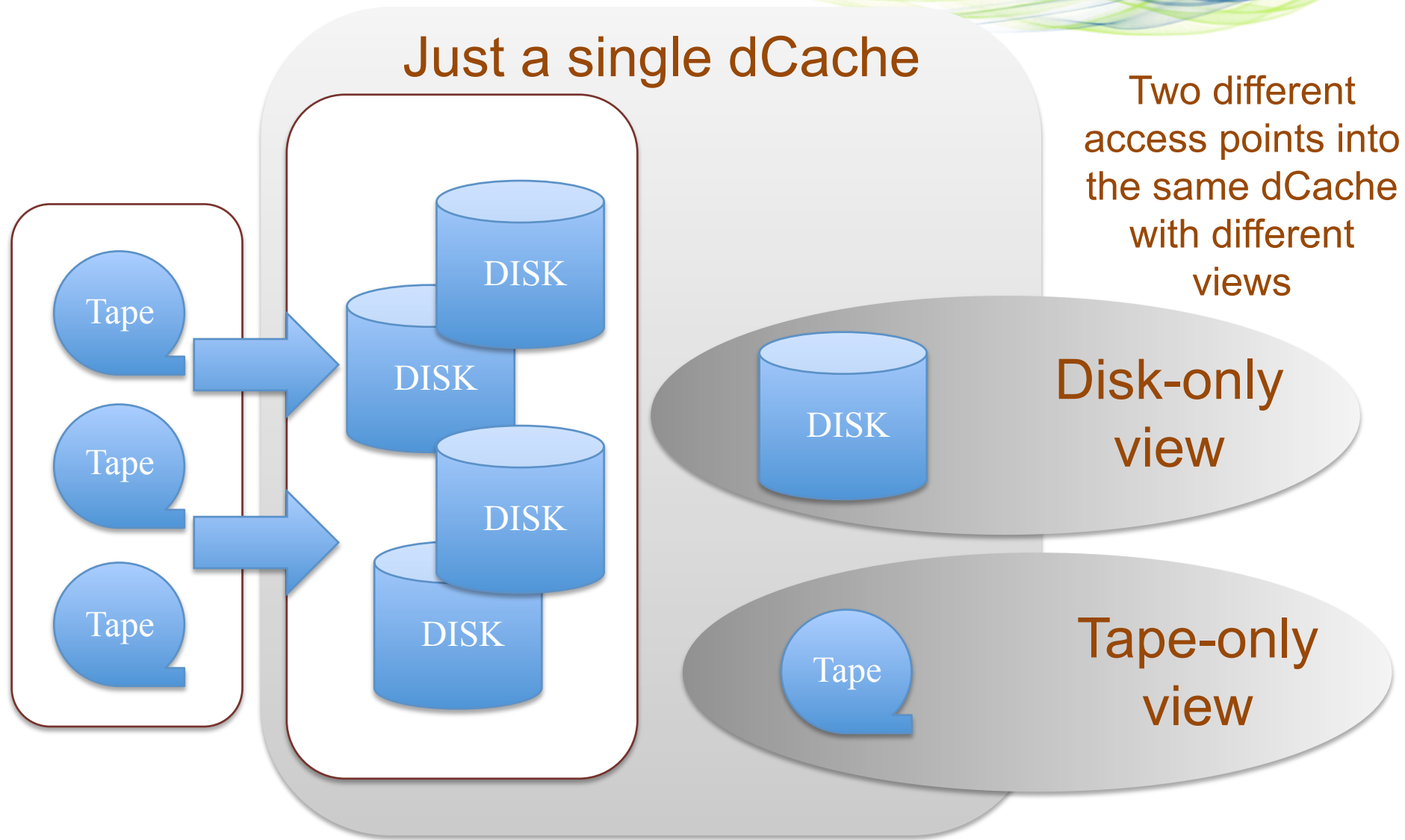
- Due to the broad developers base across international institutions and projects, dCache.org doesn't see any issues in continuous future funding.
- For the same reason, dCache.org is well integrated into the existing infrastructures and communities and keeps on track on upcoming requirements in storage management and access.
- By involving universities and students in the design and development process, dCache is keeping up with the latest developments in computers science and on the requirements of young people in data access and data sharing.



The End

further reading
www.dCache.org

CMS Disk Tape Separation





- Copy file to the tape-only part which already exists in the disk-only part.
- Remove data from disk or tape only.
- **Delayed/deferred flush to tape**
 - Very interesting for other communities.
 - Completes data lifecycle.



- People like “Storage Management” features in dCache:
 - Flush to tape
 - Recall from tape
 - Etc
- People like mounting dCache with NFS 4.1/pNFS
- People ask for (limited) “Storage Management” via NFS mounts.
 - Check File Location (done)
 - Bring Online (in preparation)

LSDMA Data Management

(More examples)



- Implementing the **CDMI Cloud protocol in dCache**, possibly followed by S3.
UNICORE will implement the client part as a proof of concept.
- Supporting the NEXUS and HDF5 file format in dCache, which essentially means **supporting non-immutable data**.
- Building an **NFS 4 German data federation**, possibly using FedFS.